

# **FIREWACH: High-throughput Functional Detection of Transcriptional Regulatory Modules in Mammalian Cells**

Matthew Murtha<sup>1</sup>, Zeynep Tokcaer-Keskin<sup>1</sup>, Zuojian Tang<sup>2</sup>, Francesco Strino<sup>3</sup>, Xi Chen<sup>4</sup>,  
Yatong Wang<sup>1</sup>, Xiangmei Xi<sup>1</sup>, Claudio Basilico<sup>1</sup>, Stuart Brown<sup>2</sup>, Richard Bonneau<sup>3,4</sup>,  
Yuval Kluger<sup>3</sup>, and Lisa Dailey<sup>1</sup>

<sup>1</sup> Department of Microbiology, NYU School of Medicine, New York, NY, 10016

<sup>2</sup>Center for Health Informatics and Bioinformatics, New York University School of  
Medicine, New York, NY 10016

<sup>3</sup> Department of Biology and Biological Sciences, Yale University, New Haven, CT  
06511

<sup>4</sup> Department of Biology, New York University, New York, NY 10003

<sup>5</sup> Courant Institute of Mathematical Sciences, Department of Computer Science, NY, NY  
10012

**corresponding author: Lisa Dailey [lisa.dailey@nyumc.org](mailto:lisa.dailey@nyumc.org)**

## **Supplementary Notes**

### **Technical details of Methods and related discussion**

#### **Table of Contents**

A. Cell culture	p3
B. Preparation of NFR-DNAs	p3
C. Estimations of the complexity of the NFR DNA populations	p4
D. Construction of Lentiviral reporter plasmid FpG5 and positive control plasmid Fgf4Enh-LV	p5
E. Preparation of NFR DNAs for cloning into LV reporter plasmids	p6
F. Cloning, bacterial transformation, and isolation of LV plasmid library DNA	p7

G. Preparation and Titre of Lentivirus	p8
H. Transduction and FACS of ESCs	p9
I. Determination of Transgene Copy Number	p10
J. Luciferase Assays	p10
K. PCR Rescue of Functionally Selected NFR-DNAs and High-Throughput Sequencing	p11
L. Genomic alignment	p12
M. Determination of Transduction Efficiency and Correlation among Replicates	p13
N. <i>In Silico</i> Generation of Random genomic DNA Fragments	p14
O. Bioinformatic and Integrative Data Analysis	p14

## Supplementary Figures

<b>Supplementary Figure 1.</b> Association of input library NFR DNAs with Genomic Regions of DNaseI Hypersensitivity.	p17
<b>Supplementary Figure 2.</b> <u>Determination of the False-Negative Rate (FNR)</u> and the percentage of GFP+ cells observed after transduction of ESCs with the Input NFR-LV Libraries.	p18
<b>Supplementary Figure 3.</b> Secondary LV Libraries Derived from FIREWACH NFRs Show Enrichment of Active CRMs.	p20
<b>Supplementary Figure 4.</b> FIREWACH Selects Elements with a Wide Range of Activity	p21
<b>Supplementary Figure 5.</b> FIREWACH Specificity	p22
<b>Supplementary Figure 6.</b> FIREWACH Sensitivity	p23
<b>Supplementary Figure 7.</b> Factors influencing the balance of False Negative and False Positive Rates in FIREWACH.	p24
<b>Supplemental Figure 8:</b> Correlation between technical and biological replicates	p25
<b>Supplementary Figure 9.</b> Comparison of Chromatin Marks Over Random, Library, and FIREWACH Elements.	p26
<b>Supplementary Figure 10.</b> Motifs Enriched in FIREWACH DNAs.	p27
<b>References</b>	p28

## Supplementary Notes

### Technical details of Methods and related discussion

#### A. Cell Culture

E14ESCs were obtained from ATCC (ES-E14TG2a, CRL-1821) and were maintained in 2i/LIF<sup>1</sup> with the inhibitors CHIR99021 (3 $\mu$ M), and PD0325901 (1  $\mu$ M) (Axon Medchem BV, The Netherlands) and 100 u/ml LIF in N2B27, or in standard ESC medium (DMEM supplemented with 15% FCS (Stem Cell Technologies, “ES cult”), 0.1mM non-essential amino acids, 0.1mM  $\beta$  mercaptoethanol, 1X Glutamax (Invitrogen), and 1000 u/ml LIF on plates coated with 0.2% Gelatin<sup>1</sup>. 3T3 were maintained in DMEM (Invitrogen), 15% Bovine Calf Serum and Penicillin/Streptomycin.

#### B. Preparation of NFR-DNAs.

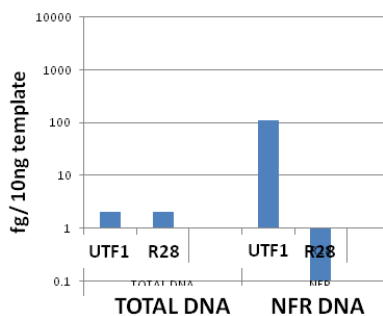
A detailed protocol for the method of extracting DNA from the NFRs of formaldehyde-crosslinked permeabilized cell nuclei has been reported <sup>2</sup>. Briefly, cultures of E14 ESCs were plated in the absence of feeder cells on eight 15 cm gelatinized tissue culture plates, and grown to 70-80% confluency, (approximately  $8 \times 10^7$  cells). The cells were crosslinked using 1% Formaldehyde in DMEM for 10 minutes at room temperature, quenched using 0.125M Glycine at RT for 10 minutes, washed with cold PBS, and collected using 2 ml/plate of PBS into 15 ml conical polyethylene tubes on ice with cell scraper. The cells were pelleted by centrifugation and stored at -80°C. Nuclei were permeabilized by resuspension in lysis buffer<sup>3</sup> and incubation for 10 minutes on ice with occasional mixing. The suspension was then dounced 10 times (B pestle), and centrifuged at 2 K rpm at 4°C for 10 minutes. The cells were resuspended in 6.4 ml Buffer 2, incubated for 10 minutes at room temperature on a platform rocker, and the nuclei were pelleted by centrifugation for 10 minutes at 2K rpm (4°C).

Pellets of permeabilized nuclei prepared from  $\sim 4 \times 10^7$  cells were resuspended in 2.6 ml of NEB2 (New England Biolabs), and distributed as five 500  $\mu$ l aliquots in eppendorf tubes on ice. 100 units of HaeIII or RsaI restriction enzyme (New England Biolabs) were each added to the NFR DNA samples, 2 tubes for each enzyme. All samples were incubated at 30°C for 1 hour with gentle mixing every 15 minutes. The reaction was

stopped with 20mM EDTA and the samples centrifuged. The supernatants were transferred to new Eppendorf tubes and re-centrifuged at maximal speed for 20 seconds. NFR-DNAs in the supernatants were either subjected directly to crosslink reversal or treated to two rounds of phenol:chloroform extraction prior to crosslink reversal.

### C. Estimations of the complexity of the NFR DNA populations.

In order to ensure that the libraries that we build will be comprehensive and representative for the full range of isolated NFR DNAs, it is necessary to be able to estimate the complexity, i.e. the number of unique elements, present in the NFR populations. To this end, a quantitative PCR (qPCR) approach was implemented (Primer sequences in Supplementary Table 4). An approximately 200 bp PCR product encompassing the UTF1 enhancer was generated using genomic mouse DNA as template, and gel purified. The purified UTF1 DNA was quantified and diluted in 10-fold intervals over a range of 1fg to 1 ng. The UTF1 DNAs within this dilution series were used as templates in qPCR for creating a standard curve, allowing each quantity of UTF1 DNA to be correlated with a specific number of amplification cycles. The same set of PCR primers was used to amplify UTF1 enhancer DNA sequences from 10ng of the HaeIII- or RsaI- NFR DNA preparations. The number of cycles needed to detect UTF1 enhancer DNA in each of the NFR samples was compared to the standard curve, assayed in parallel, permitting an estimation of the amount of NFR DNA that corresponds to the UTF1 enhancer in each preparation. For example, 10ng of the HaeIII NFR preparation contains approximately 100fg of UTF1 enhancer DNA. Our previous work showed that NFR DNAs isolated by our method are small, with an average length of 150 bp. By making the assumption that the NFR fragments within the total population are of a similar length as the UTF1 enhancer fragment, the results of the qPCR analysis would indicate that the UTF1 enhancer DNA corresponds to approximately 1 out of every  $10^5$  fragments in the HaeIII NFR population (Figure 1). Thus the complexity of this population is estimated to be  $10^5$ .



**Figure 1. qPCR analysis to estimate the complexity of NFR DNA populations.** 10ng of Total genomic DNA, or LMPCR amplified HaeIII-NFR was used as template in qPCR reactions containing primers for the detection of UTF1

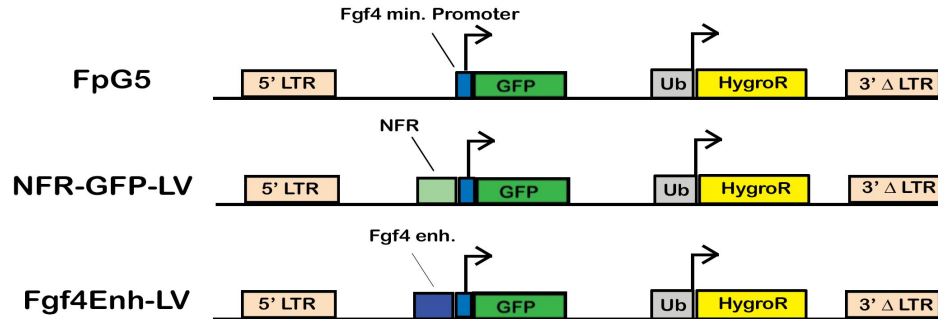
enhancer or R 28 DNA. The number of cycles required to detect these DNAs in each of the DNA samples was compared to standard curves for UTF1 and R28, as described in the text, to determine the relative amount of each of these sequences.

A standard curve was similarly created for R28 DNA and the analysis of R28 DNA in the NFR populations using qPCR confirmed that R28 DNA is detected at the same level as UTF1 DNA from the total genomic DNA template, but was undetectable in the NFR DNA samples (Fig.1). Thus in addition to providing us with an assessment of NFR DNA complexity, these results present a clear, quantitative picture of the selectivity of our method for the isolation of accessible DNA regions from ES cell chromatin.

#### **D. Construction of Lentiviral reporter plasmid FpG5 and positive control plasmid Fgf4Enh-LV.**

The DNA plasmid for generation of the self-inactivating lentivirus FUW was obtained from Addgene (Addgene # 14882)<sup>4</sup>. Coding sequences of the Hygromycin resistance gene were generated using PCR amplification of plasmid pCEP4 using primers PR12 and PR13 (Supplementary Table 4). The amplification product, containing BglII sites at each end, was inserted at the unique BamHI site immediately downstream of Ubiquitin promoter DNA sequences within FUW, creating a hybrid BamHI/BglII site that is resistant to digestion by either enzyme. The resulting construct was named FUWH. A DNA cassette containing the Fgf4minimal promoter upstream of GFP coding sequences and transcription stop/polyA signals was PCR amplified using primers PR10 and PR11 (Supplementary Table 4) with the -64GFP plasmid DNA as template<sup>5</sup>. In parallel, a cassette containing DNA sequences of the Fgf4 enhancer, Fgf4 minimal promoter, and GFP coding sequences and transcription stop/polyA signals was PCR amplified using oligonucleotide primer sequences PR14 and PR11 (Supplementary Table 4) with enhGFP plasmid DNA as template<sup>5</sup>. Due to the design of the primer sequences, these amplification products contain a PacI recognition site at both the 5' and 3' ends. Thus the promoter-GFP and enhancer-promoter-GFP cassettes were each cloned, in both orientations, into the single PacI site upstream of the Ubiquitin promoter within FUWH. Assessment of GFP expression following the transduction of these lentiviruses into F9 cells indicated that plasmids containing the colinear orientation of the GFP and Hygromycin units resulted in somewhat better GFP expression (data not shown). The promoter-GFP lentiviral construct, FpG5, and the enhancer-promoter-GFP lentiviral

construct, Fgf4Enh-LV, are depicted in Figure 1. *FpG5* has BamHI site proximal to the promoter that accepts BglII digested NFR/Adaptor DNA oligos.



**Figure 2.** GFP-LV reporter constructs. FpG5 contains the minimal promoter (TATA box and TSS of the Fgf4 gene) and GFP coding sequences within a LV plasmid that also contains a hygromycin resistance gene driven by the Ubiquitin promoter. NFR DNAs derived from mouse ESCs (red box, NFR-GFP-LV) are cloned immediately upstream of the minimal promoter to generate the NFR-GFP-LV libraries. Fgf4Enh-LV is a positive control LV construct containing 150bp segment of the well-characterized ESC-specific Fgf4 enhancer (blue box) upstream of the fgf4 promoter-GFP cassette.

#### E. Preparation of NFR DNAs for cloning into LV reporter plasmids

The protocol for the preparation of the double-stranded adaptor DNAs and their ligation to NFR-DNAs is detailed elsewhere <sup>2</sup>. Briefly, HaeIII NFR-DNAs and RsaI NFR-DNAs were each subjected to blunt-end ligation to distinct adaptor DNAs that, after ligation, permit the restoration of the HaeIII- or RsaI sites, respectively, at the NFR/Adaptor junction (Supplementary Table 4). The HaeIII and RsaI adaptors were generated by annealing equimolar amounts of the respective “Linker A” and “Linker B” oligos (Supplementary Table 4). The adaptor sequence contains a BglII site that used for cloning. Annealed linkers were ligated to NFR DNAs overnight at 16°C using the following reaction:

30 µl NFR-DNA supernatant, 10 µl 5X Ligation Buffer (Invitrogen), 6.7 µl 15 uM annealed Linker, 2.3 µl H<sub>2</sub>O, 1 µl (5 Units) T4 DNA Ligase (Invitrogen).

After ligation, the DNAs were purified using the QiagenMini-Elute PCR cleanup kit and eluted in 30 µl H<sub>2</sub>O. 25 µl of the eluted DNA were assembled in a 50 µl reaction for PCR

amplification using the appropriate “AMP” oligonucleotide primer (Supplementary Table 4) and the following PCR conditions: 55°C for 2 min to melt the shorter “Linker B” oligo away from the NFR/Adaptor and then 72°C for 5 min, 95°C for two min, followed by fifteen cycles of 95°C 1', 60°C 1', and a final extension at 72°C for 5 min. The amplified DNAs were purified using the Qiagen PCR cleanup kit and eluted in 50 µl H<sub>2</sub>O.

#### **F. Cloning, bacterial transformation, and isolation of LV plasmid library DNA.**

Purified PCR-amplified NFR DNAs were digested with BglII restriction enzyme digestion overnight and purified using a Qiagen PCR cleanup column using 50 µl water for elution. The DNA concentration of the samples was determined using Nanodrop. Multiple ligation reactions were assembled containing 200 ng BamHI-digested and phosphatase-treated FpG5 LV vector plus 40 ng of BglII-cut, LMPCR-amplified NFR DNAs in a 20 µl total ligation reaction using 5 u T4 DNA ligase (Invitrogen, or Roche). Ligation was performed overnight at 16°C.

Commercially available electro-competent Stbl4 bacteria were used for high efficiency transformation using electroporation according to the parameters suggested by the manufacturer (Invitrogen). After purification of the ligation reactions through QiagenMini-Elute columns and elution in 20 µl of water, 1-2 µl of each reaction were used for the electroporation of 20 µl Stbl4. 700-900 S.O.C. broth was added to the Electroporated cells and, after 1 hour recovery, the sample was divided in three and spread over 3 15 cm Agar plates containing 50 µg/ml Ampicillin. This procedure generally yielded several thousand colonies per electroporated sample.

The ligation efficiency for each reaction was determined by transferring cells from 20 colonies into tubes containing PCR reaction components and primers PR2 and PR21 that are complementary to sequences flanking the LV BamHI cloning site. PCR amplification was performed for 25 cycles (95°C 30", 58°C 1 min, 72°C 30") and run in a 2% agarose gel in TAE buffer. The percentage of constructs containing insert was determined by the presence of a PCR product migrating slower than that amplified from FpG5 template. Generally, 80-90% of the constructs contained an NFR DNA insert.

These steps were repeated until approximately  $5 \times 10^5$  colonies each for the HaeIII and RsaI constructs were obtained. The ampicillin plates were stored at 4°C until collection.

To prepare the library DNA, 5 ml of cold LB containing 10% Glycerol were added to the plate, and the colonies were collected without further amplification using a cell scraper across the plate surface. All colonies were collected into a single flask, mixed, and then divided into 2 portions. One tube was stored at  $-20^{\circ}\text{C}$  and the other half used for preparing library DNAs.

To prepare library DNA from the pooled colonies, we used a Qiagen maxiprep kit following standard protocols for plasmid DNA isolation.

### **G. Preparation and Titre of Lentivirus**

NFR-lentiviral libraries were prepared using ViraPower (Invitrogen) following manufacturer's protocol. Briefly,  $5 \times 10^6$  293FT cells were plated to a p-lysene coated 10 cm dish one day prior to transfection in 10% FCS DMEM without antibiotic. On the day of transfection, the medium was replaced with 5 ml of Opti-DMEM/10% FCS without antibiotic. For each NFR-lentiviral library and control lentivirus DNA-Lipofectamine 2000 complexes were generated as follows. 9 ug of ViraPower Packaging Mix and 3 ug of lentiviral plasmid were diluted into 1.5 ml of Opti-MEM medium without serum. In a separate tube 36 ul of Lipofectamine 2000 was diluted in 1.5 ml Opti-MEM and allowed to incubate at RT for 5 min. After incubation diluted DNA and Lipofectamine 2000 solutions were combined with gentle mixing and allowed to incubate at RT for 20 min. The solution was then added dropwise to a single 10 cm dish of near confluent 293T cells and incubated overnight at  $37^{\circ}\text{C}$ . The next day medium was changed to complete ESC medium without LIF. Virus containing media was collected at 48 and 72hrs post transfection and stored at  $-80^{\circ}\text{C}$ .

Prior to freezing lentiviral titres were determined using p24 Antigen ELISA (ZeptoMetrix). Virus containing media was diluted  $10^3$  and  $10^4$  fold with DMEM in 450  $\mu\text{l}$  aliquots. 50  $\mu\text{l}$  of lysis buffer was added to each sample. A six point p24 antigen standard curve was generated by successively diluting 125 pg/ml solution of p24 antigen 1:2 to a final concentration of 7.8 pg/ml. 200 ul of standard sample or lentiviral containing media was added to individual wells of the p24 ELISA microplate, covered with plate sealer, and allowed to incubate for 2hr at  $37^{\circ}\text{C}$ . After incubation the wells were aspirated and washed five times with 300  $\mu\text{l}$  wash buffer. 100  $\mu\text{l}$  of HIV-1 p24 Detector Antibody is then added to each and incubated at  $37^{\circ}\text{C}$  for 1 hr. Wells are washed as before and 100

μl streptavidin-peroxidase working solution is added to each well incubated for 30 min at 37°C. Wells are washed and 100 μl freshly prepared Substrate Working Solution is added to all wells and incubated uncovered at room temp for 15 minutes. 100 μl of Stop Solution is then added and the optical density, OD, of each well is immediately measured at 450 nm using a Spectromax M5 plate reader. The slope, b, and intercept, m, of the standard curve is determined and the final concentration of lentiviral particles per sample is inferred with the equation [Titer = (OD-b-blank)/m x 100 x dilution factor].

#### H. Transduction and FACS of ESCs

ESCs were transduced at a MOI of 7 to ensure that the maximum number of cells is transduced while favoring single copy integration. To increase the likelihood that any given NFR-lentiviral genome would be represented, the number of cells transduced was equivalent to more than ten times each library's complexity. Thus  $5 \times 10^6$  E14 ESCs were plated in complete ESC medium plus LIF in feeder free conditions on a 10 cm gelatin coated dish one day prior to transduction. The cells were then transduced overnight in 10 ml of complete ESC medium plus 8 μg/ml polybrene, containing  $3.5 \times 10^7$  virus particles for a MOI of 7. The following day the medium was replaced with fresh ESC medium plus LIF. Hygromycin-selection was initiated four days post transduction in ESC medium/LIF containing 250 μg/ml hygromycin B. Cells were selected for hygromycin B resistance for 5 days, with media changed daily. GFP+ cells were selected using fluorescence activated cell sorting (FACS) on a iCyt Reflection HAPS2 cell sorter. Cells were treated with propidium iodide (2 μg/ml) prior to sorting to counter-select dead cells. The gate was set relative to the profile of FpG5 transduced cells such that the number GFP+ cells observed was less than 0.5%. Cells transduced by NFR-GFP-LV and expressing GFP at a level higher than this set point were collected using FACS. Collected cells were returned to culture, expanded, and subjected to additional 1-2 rounds of FACS to obtain a population of greater than 90% GFP+ cells. A minimum of  $10^6$  GFP-positive cells was collected from each sort so as to maintain complexity of the integrated transgene population. Post-sort FACS analysis was performed with a minimum  $10^5$  cells per 100 μl sort buffer on a FACSCalibur flow cytometer (BD Biosciences) and analyzed with FloJo software.

For each NFR-GFP-LV library, i.e. HaeIII and RsaI derived libraries, two independent transductions were performed to generate two biological replicates for each library. Each

replicate was transduced, selected for hygromycin resistance, and sorted independently to generate cell lines (HaeIII\_BioRep1, RsaI\_BioRep2, etc.) comprised of pools of NFR-GFP-LV transduced cells. Each cell line was cultured and independently assayed for copy number and NFR sequences. Downstream informatics analysis was also largely done on independent lines prior to pooling end-result NFR sequence information and analysis.

## **I. Determination of Transgene Copy Number**

Average copy number of integrated lentivirus was estimated using an adapted qPCR approach <sup>6</sup>. Briefly, genomic DNA from each transduced cell line was obtained from  $1 \times 10^6$  cells with DNeasy (Qiagen). The number of lentiviral vector genomes per cell was determined by quantitative real-time PCR with primers recognizing the GFP transgene while number of mouse genomes was determined using primers recognizing a unique noncoding region of the genome (Primers “Gen-F” and “Gen-R”). A six point standard curve from  $1^8$  to  $1^2$  copies was generated by serial dilution of a single plasmid cloned to contain both the GFP and genomic DNA target elements. Amplification reactions contained 5  $\mu$ l Sybergreen MasterMix, 2  $\mu$ l gDNA (100 ng), 2  $\mu$ l H<sub>2</sub>O, and 0.5  $\mu$ l each of 5  $\mu$ M forward and reverse primer. Reactions consisted of 40 cycles at 95°C (15s) then 60°C (1 min) on a BioRad thermocycler. Data were plotted against and interpreted in the linear portion of the standard curve where regression coefficient was greater than 0.98. The average integrated copy number was determined by dividing the calculated number of lentiviral genomes by the total number of mouse genomes present in the DNA sample of each transduced line and measured in triplicates.

## **J. Luciferase Assays**

### **a. Reporter constructs**

The pGL3 luciferase reporter plasmid was modified to contain the 162bp minimal Fgf4promoter (fgfprom-luc). This plasmid has a BglII site upstream of the fgf4 promoter sequences used for inserting test DNAs. Oligonucleotide primers used to recover library NFR or FIREWACH DNAs from the lentiviral plasmids and prepare them for InFusion cloning into the fgfprom-luc plasmid were designed as follows: the 5' portion consisted of 15 bases complementary to the sequence flanking the fgf4prom-luc plasmid BglII site, and the 3' portion contained sequences complementary to sequences flanking the NFR DNA cloning site within the lentiviral plasmid. PCR amplification was performed using

either input lentiviral NFRGFP-library plasmid DNA or the genomic DNA isolated from FACS-sorted GFP<sup>+</sup> cells as template. The amplified fragments were used for In-Fusion (Clontech) directional cloning into the fgfprom-luc plasmid. Primers InFusionpGL3R, InFusResFA were used to clone into the proximal BglII site of fgfprom-luc while DisInFusResFA and DisInFuspGL3R were used at the distal BamHI site of TKluc (Supplementary Table 4). Recombinase reactions were assembled according to the manufacture's protocol.

To generate luciferase reporter constructs to assay random genomic DNA fragments, three micrograms of purified gDNA were digested with either HaeIII or RsaI, and DNA fragments ranging from 100-300bp were gel-purified and cloned into the SmaI site of the fgfprom-luc plasmid.

#### **b. Transfections and luciferase assays**

E14 cells grown in ESC medium with 1000U/ml LIF were seeded on 0.2% gelatin coated 96 well plates at  $5 \times 10^4$  cells/well. Cells were transfected using 250ng plasmid DNA and 1.25ul Lipofectamine 2000 (Invitrogen) and supplemented with OPTIMEM and LIF (1000 u/ml) for a total volume of 150 ul/well. 4 hours after transfection the medium was changed to complete ESC medium plus 1000U/ml LIF. 24 hours after transfection, lysates were prepared and luciferase assays were performed as instructed by the manufacturer (Promega). The protein concentration of the lysates was determined (Bio-Rad) and used to normalize the samples. The luciferase activities of all test constructs were calculated relative to the activity displayed by the fgfprom luciferase construct containing only the minimal fgf4 promoter upstream of the luciferase gene.

#### **K. PCR Rescue of Functionally Selected NFR-DNAs and High-Throughput Sequencing**

NFR-DNAs were rescued from either the initial lentiviral plasmid libraries or gDNA of GFP<sup>+</sup> selected cells using PCR in a method adapted from bacterial rRNA sequencing<sup>7</sup>. In this method, Illumina sequencing adaptors are included in the primers, permitting one step amplification and sequencing library preparation<sup>7</sup>. Primers (termed FMS-F/R, Supplementary Table 4) were designed such that they contain recognition sequences complementary to lentiviral sequence flanking NFR-DNA and Illumina adaptor sequence for paired-end sequencing in addition to a 6base pair Index sequence. Six PCR

reactions (10 µl Phusion Polymerase buffer, 1 µl 10 mM dNTP, 2.5 µl 10 µM forward and reverse primer, 1.5 µl DMSO, 0.5 µl (NEB) 50 ng DNA, 31 µl H<sub>2</sub>O; 16 cycles with 55°C annealing temperature) per plasmid library were pooled and sequenced.

FIREWACH elements were recovered from the genomic DNA of a least 1x10<sup>6</sup> FACS-sorted GFP+ cells using PCR. In this case 10 PCR reactions were performed using the same conditions as above but 100 ng gDNA and 23 cycles of amplification. The 10 reactions were then pooled.

Each sample was amplified with primers containing Illumina adaptor sequence with 6bp indexing sequence. This allowed us to pool up to six samples within a single lane on the MiSeq machine. Input library derived NFRs were sequenced together using three of the barcodes while FIREWACH NFRs, i.e. NFR's rescued from GFP+ cells, were run with six samples per lane, each sample representing NFRs rescued from an independent biological replicate (i.e. HaeIII\_BioRep1 etc). Technical replicates consisting of independent PCR rescued NFR sequencing libraries were sequenced on separate days.

Samples were run on a miSeq sequencer with the miSeq cartridge version 2, as a 2 X 150 bases run, with a 50% PhiX library spiked in to compensate for potential low diversity in the libraries. In order to ensure efficient binding of sequencing primers we designed and used custom Read 1, Index and Read 2 primers (sequences in Supplementary Table 5) of which 17 µl of custom primers at 100 µM were spiked into the Illumina Read 1, Read 2 and index reads positions in the cartridge.

#### **L. Genomic alignment**

Illumina MiSeq 2x 151 bp data were pre-processed by demultiplexing and trimming of 7 bp from the 5' end and 44 bp from the 3' end, yielding a data set of 2 x 100 bp sequence reads. Paired end sequences were aligned to the mouse reference genome (mm9) using BWA<sup>8,9</sup> software with default settings. Read pairs were filtered from the final data set if either read failed to map to the genome, if both reads did not map in the proper orientation, if the mapping quality score of both reads was less than 25, or if neither read had a unique map location on the genome. Target sites were identified as loci where paired reads both aligned entirely within a 500 bp genomic region.

Each Biological Replicate sample was independently sequenced three times (i.e. three technical replicates) and all sequencing data for all samples were then merged to create the final list of FIREWACH genomic regions (6,364 elements). The final input NFR DNA library dataset was generated by merging the mapped loci from replicate of each enzyme library as well as with all FIREWACH loci generating a list of 84,240 elements.

#### **M. Determination of Transduction Efficiency and Correlation among Replicates**

Transduction efficiency was estimated by PCR rescue and sequencing of a portion of HaeIII\_BioRep2 prior to FACS. This sampling yielded 3,238 genomic loci, 30% of which (i.e. 995 loci) corresponded to elements on the FIREWACH list. That 30% of the transduced elements correspond to predicted active elements is in overall agreement with the percentage of individually tested library DNAs that were observed to activate expression in the luciferase assays of Supplementary Fig. 5. The 995 FIREWACH loci identified within the transduced cell sampling represent 20% of the total number of HaeIII NFR DNAs present in the final FIREWACH list, suggesting that a minimum of 5 times the number of constructs would have had to be transduced into ESCs to achieve the final number of FIREWACH elements. Additional consideration of the False Negative rate of 0.26 (Supplemental Fig. 2) further indicates that transduction of at least 1.35 this number would be required for the activity of all FIREWACH elements to be detected. Thus the minimal number of input library constructs transduced is estimated to be 21,856.5 (i.e.  $3,238 \times 5 \times 1.35$ ), or 56% of the total input HaeIII-GFP-LV library.

The correlation coefficients for technical or biological replicates were calculated by binning the genome into windows of 100bp and computing the Pearson correlation of the genomic coverage between all pairs of the coverage vectors, which represent our sequencing datasets. The calculation was done using an in-house Java code. Technical replicates consist of independent sequencing library preparations from a common template (e.g. GFP+ cells transduced with HaeIII NFR-GFP LV), and were generated to assess the reproducibility of our sequencing library preparation protocol. Three independent runs of each biological replicate were compared pair-wise and the average of all taken for a given enzyme-derived NFR library (eg. 0.86-0.98 for HaeIII\_BioRep1). These replicates did not correlate with random NFRs generated in silico (Average of 0.001 for HaeIII and RsaI both).

For measuring the correlation between biological replicates, the total reads from all three technical replicates of FIREWACH-seq elements were combined into a single file. For example, HaeIII\_BioRep1, contains three technical sequencing replicates generated from the recovery of cloned NFR DNAs from the integrated LV vectors within HaeIII\_BioRep1 transduced GFP+ ESC. The HaeIII\_BioRep1 sequences were combined with RsaI\_BioRep1 sequences into a single file (Rep1). Rep2 was similarly generated from HaeIII\_BioRep2 and RsaI\_BioRep2. Comparison of Rep1 and Rep2 generated the correlation between biological replicates (0.61).

#### **N. *In Silico* Generation of Random genomic DNA Fragments**

To create a dataset of random genomic DNA fragments, we generated a list of genomic loci corresponding to digestion of the murine reference genome (mm9) with HaeIII or RsaI. We utilized a script to scan chr19 for pairs of each restriction sites as a regular expression separated by a variable region of DNA up to 500bp in length so that the size distribution of the *in silico* fragments would be comparable to that of the enzymatically derived NFR libraries. The resulting *in silico* DNA fragment dataset was comparable in number to the input NFR library (61,844 random elements versus 84,240 NFR-DNAs) and was then reformatted as genomic loci in a bed file. Random distribution of elements was confirmed in subsequent analysis as this list generated correlative scores expected of a randomly distributed set of loci (e.g. in comparison to DNaseI-HS, the random elements had an AUROC=0.52, typical of random elements, Supplementary Figure 1).

#### **O. Bioinformatic Data Analysis**

To investigate the chromatin status of FIREWACH elements we utilized several publically available sequence files for ChIP-seq and DNaseI-HS sequencing experiments. Data were obtained from the NIH's sequence read archive (SRA) and the UCSC genome browser (for full list of data sets used, see Supplementary Table 5). In the case of the H3K4me1/3, H3K27me3/Ac, H3K9Ac chromatin marks, reads were remapped using the mm9 genome as reference with bowtie<sup>10</sup> (version 2) with the options “-n 1 -k 1 -m 20 --best --strata -p 8 --chunkmbs 1024”. Tophat (version 2.0.4,) and cufflinks (version 2.0.2,) with default parameters were used to obtain FPKM values for all genes<sup>11</sup> from RNA-seq data<sup>12</sup>.

##### **a. GREAT Analysis**

Bed files of all input library NFR DNAs, HaeIII library elements, or RsaI elements, or FIREWACH DNAs were analyzed using the Genomic Regions Enrichment of Annotations Tool<sup>13</sup> (<http://bejerano.stanford.edu/great/public/html/>) with the settings: Mouse: NCBI build 37 ([UCSC mm9, Jul/2007](http://ucsc.genome.ucsf.edu/mm9/)), Whole genome background, basal plus extension, Proximal 5kb upstream, 1kb downstream, plus distal up to 100kb. Datasets were analyzed using both the Significance by Both and Significance by Region-based Binomial views.

#### **b. Comparison with genomic regions of DNaseI Hypersensitivity in ESCs**

The relation between DNase I HyperSensitivity and the input library NFR DNAs was investigated using Receiver Operating Characteristic (ROC) curves. In particular, we verified that open regions (i.e. with high DNaseI HS coverage) could predict the location of genomic regions covered by input NFR DNA reads. For each dataset (All library, HaeIIINFR DNA library, RsaINFR DNA library and *in silico* DNA library), the genome was divided into non overlapping bins of 1kbp and the bin was classified as positive if it contained at least one NFR DNA read or negative if it did not intersect any element. The coverage of DNase I Hypersensitivity HotSpots (ES-CJ7 Pk1, UCSC genome browser) for each bin were used as classifier in order to build the ROC curve (i.e. DNase I HS coverage is utilized to predict whether a bin would contain any of the elements). The area under the curve of the receiver operating characteristic (AUCROC) was 0.8637 for the combined (All) library, 0.8780 for RsaI-, and 0.8539 for HaeIII DNAs. The AUCROC for the *in silico* reads (0.5267) was not significantly different from 0.5, which is the expected value of random reads. The area under the curve was calculated and plotted in graph form as presented in Supplementary Figure 1.

The 84,240 input NFR library DNAs comprise a total 4,555,888 bps, which corresponds to approximately 4% of the 113,439,159 bps of DNA contained within the regions of DNase Hot Spots in ESCs.

#### **c. RNA-seq Analysis**

For each unique proximal read, we considered the expression of the nearest gene. The expression data for ES cells in 2i medium were obtained from Marks *et al.* Tophat (version 2.0.4)<sup>11</sup> and cufflinks (version 2.0.2)<sup>11</sup> with default parameters were used to obtain FPKM (Fragments Per Kilobase of transcript per Million mapped reads) values for

all genes. To aid visualization and analysis, we scaled the FPKM values logarithmically as  $\log_2(1+\text{FPKM})$ . The significance between different libraries was assessed using the nonparametric Kruskal-Wallis test<sup>14</sup>.

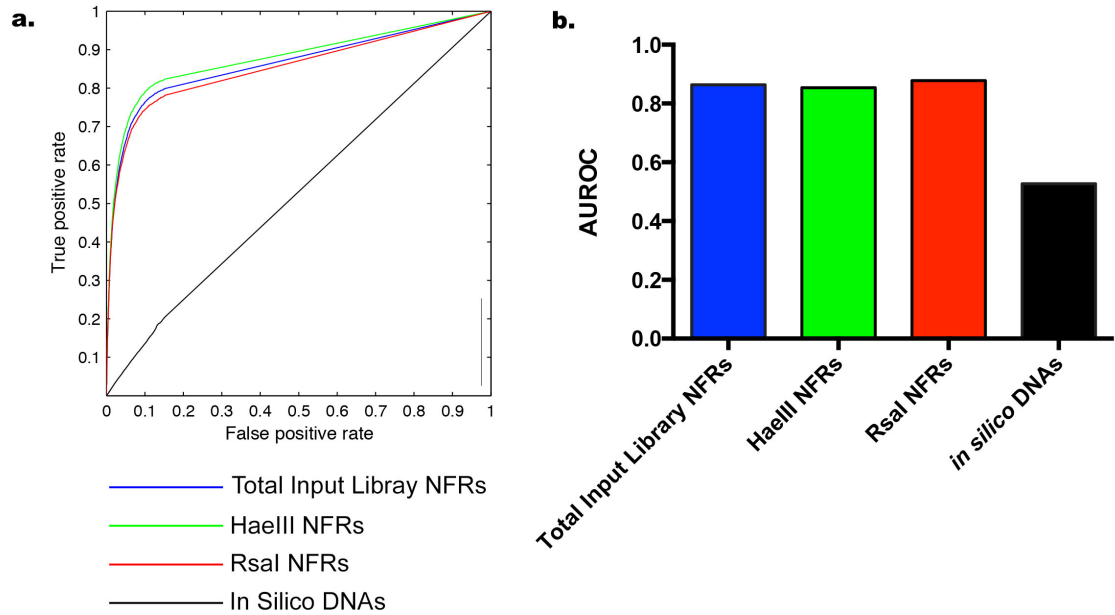
#### **d. Carpet methods**

High-density maps of coverage of chromatin marks around FIREWACH loci was visualized as previously described<sup>15,16</sup>. Each horizontal line represents the center of a unique FIREWACH DNA. The expression of the nearest gene is color-coded from red (expressed) to green (not expressed) and the expression values are used to sort the horizontal values. The ChIP-seq signal in the  $\pm 1$  kb region around each FIREWACH locus was determined for H3K4me1 H3K4me3, H3K27me3, H3K27Ac, and DNaseI hypersensitivity. The ChIP-seq and DNaseI HS signals were normalized by total number of reads, The gene expression data was quantile-normalized over all genes. In case of identical, overlapping or nearby FIREWACH loci (< 100 b), the profile of only one read was used in the high-density map.

#### **e. Motif Analysis**

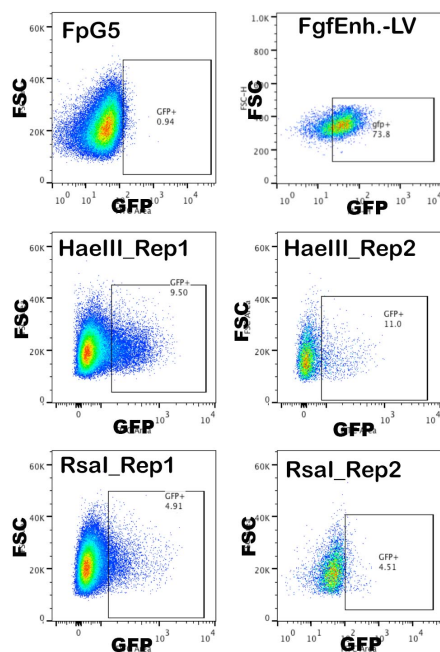
Motif enrichment analysis was performed for the distal FIREWACH elements using the AME module in the MEME suite<sup>17,18</sup> with the following command line options: “--method mhg --scoring totalhits --length-correction”. Random gDNA elements from *in silico*/ digestions were used as background model. We also analyzed distal elements using input NFRs as background. This allowed us to determine which motifs were enriched above those obtained from open chromatin alone. P-values were calculated based on the multi-hypergeometric distribution and corrected for multiple hypothesis testing. Analysis was performed using a database of known motifs that covers approximately 50% of mouse TFs (Supplementary Fig. 10 and Supplementary Table 7). This curated compendium of motifs can be accessed via the Timothy Hughes lab webpage (<http://cisbp2.ccbr.utoronto.ca/>) and is derived from protein binding microarray data, HT-SELEX, and ChIP-seq<sup>19-21</sup>. Most of the motifs used are also redundantly available in the JASPER and TRANSFAC databases<sup>22,23</sup>.

#### **Supplemental Figures.**



**Supplementary Figure 1. Association of input NFR library DNAs with Genomic Regions of DNaseI Hypersensitivity.**(a) ROC curves were calculated in order to characterize the degree of correlation between Input Library NFR DNAs and genomic regions of DNaseI hypersensitivity. The Figure shows the results of the separate analysis of the HaeIII or RsaI NFR DNAs as well as the total NFR-GFP-LV library NFRs in which the HaeIII- and RsaI DNAs have been combined and analyzed as a single sample. Each of these samples display a high degree of overlap with accessible regions as defined using DNaseI (AUROC=0.85-0.87). In contrast, similar analysis of a random set DNA fragments generated by in silico digestion of murine genomic DNA with RsaI and HaeIII generates AUROC of 0.52. **(b)** Histogram of the AUROCs from (a).

a.



b.

Library	Expected %GFP+	Observed % GFP+	FPR	FNR
FpG 5	0	0.1	0.001	-
FgfEnh-LV	100	73.8	-	0.262

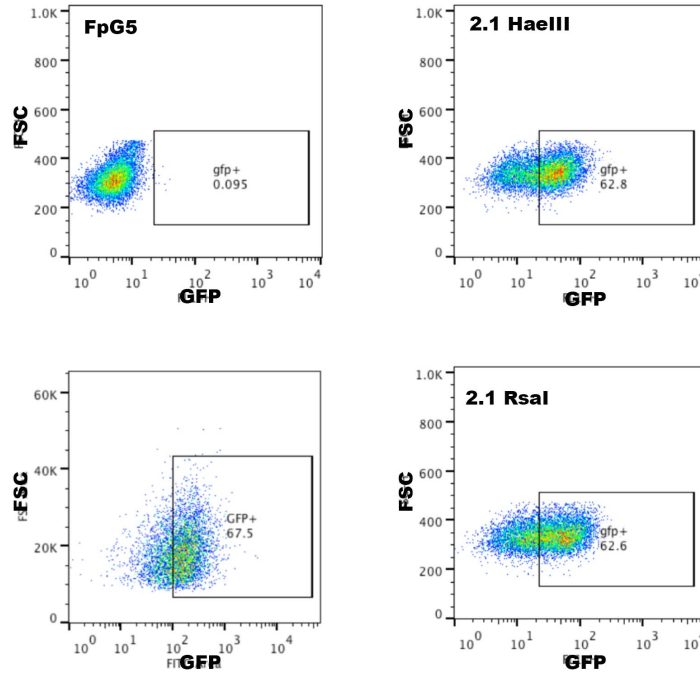
c.

Library	%GFP+
HaeIII_BioRep1	9.5
HaeIII_BioRep2	11.0
RsaI_BioRep1	4.9
RsaI_BioRep2	4.5

**Supplementary Figure 2. Determination of the False-Negative Rate (FNR) and the percentage of GFP+ cells observed after transduction of ESCs with the Input NFR-LV Libraries.** (a) Quantitative flow cytometry was used to determine the percentage of transduced cells expressing GFP after hygromycin selection. Negative control FpG5-transduced cells display virtually no GFP expression and were used to set the gate (boxed area) for detecting GFP expression in all other samples. (b) The positive control lentivirus FgfEnhLV (Figure 3) was used to calculate the False-Negative Rate. GFP gene expression from this construct is controlled by the well characterized Fgf4 enhancer that is highly active specifically in ESCs<sup>24</sup>. Although all cells harboring FgfEnh-LV should express GFP, only 73.8 % of transduced cells were observed to do so. Presumably this results from the fact that the lentiviral DNAs integrate randomly throughout the transduced ESC genome and that integration into some loci has a negative effect on transgene expression<sup>25</sup> the estimated False Negative Rate (FNR) for

active lentiviral constructs transduced into ESCs in our experiments is 0.26. Note that the opposite effect, false activation of negative elements, is not observed as evidenced by the lack of GFP+ cells in FpG5 transduced cells. In consideration of the FNR, we attempt to minimize the exclusion of true active elements from our final set of FIREWACH elements by transducing ESCs using at least a ten-fold excess over the estimated number of elements present in the input NFR-GFP- LV library so that each construct is provided a better opportunity to integrate in a permissive genomic locus. **(c)** The percentage GFP+ cells in ESCs transduced with each of the indicated NFR-GFP-LV libraries of constructs was determined using flow cytometry. BioRep1 and BioRep2 are Biological replicate samples resulting from two independent transductions for each NFR-GFP-LV library.

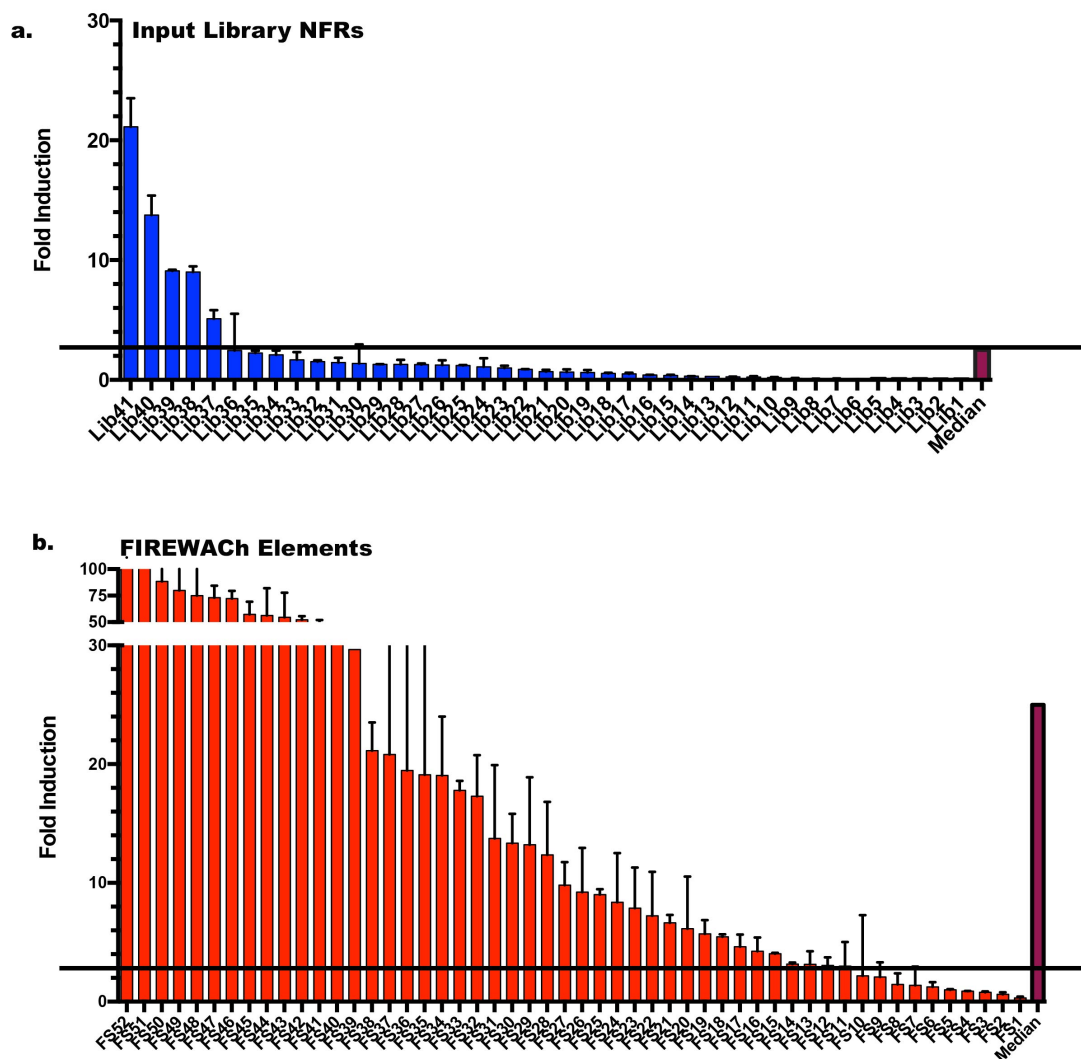
**a.**



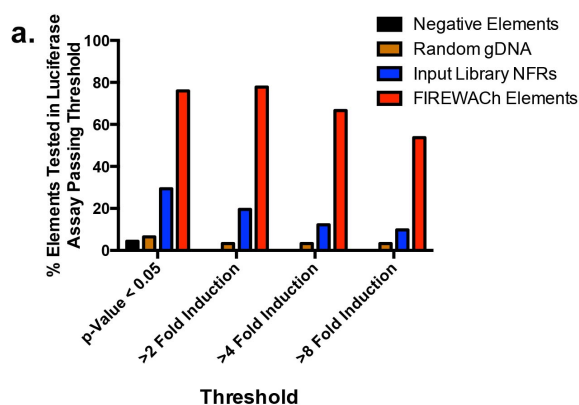
**b.**

Library	Primary %GFP+ (average)	Secondary % GFP+
HaeIII	10.3	62.8
RsaI	4.7	62.8

**Supplementary Figure 3. Secondary LV Libraries Derived from FIREWACH DNAs Show Enrichment of Active CRMs.** FIREWACH elements were recovered from FACS-purified GFP+ cells for each of the Biological replicates of HaeIII- and RsaI-NFR-GFP LV- transduced cells using PCR. Recovered DNAs from each set of Biological replicates were pooled and re-cloned into the FpG5 LV vector to create two secondary NFR-GFP-LV libraries. **(a)** Plots of quantitative flow cytometry analysis shows that ~63% of ESCs transduced by the secondary HaeIII- or RsaI- NFR-GFP-LV libraries are GFP+ **(b)** Comparison of the percentage of GFP-positive cells observed after transduction with primary or secondary LV libraries shows enrichment for active elements in the secondary libraries, demonstrating the ability of FIREWACH to functionally select active CRMs.



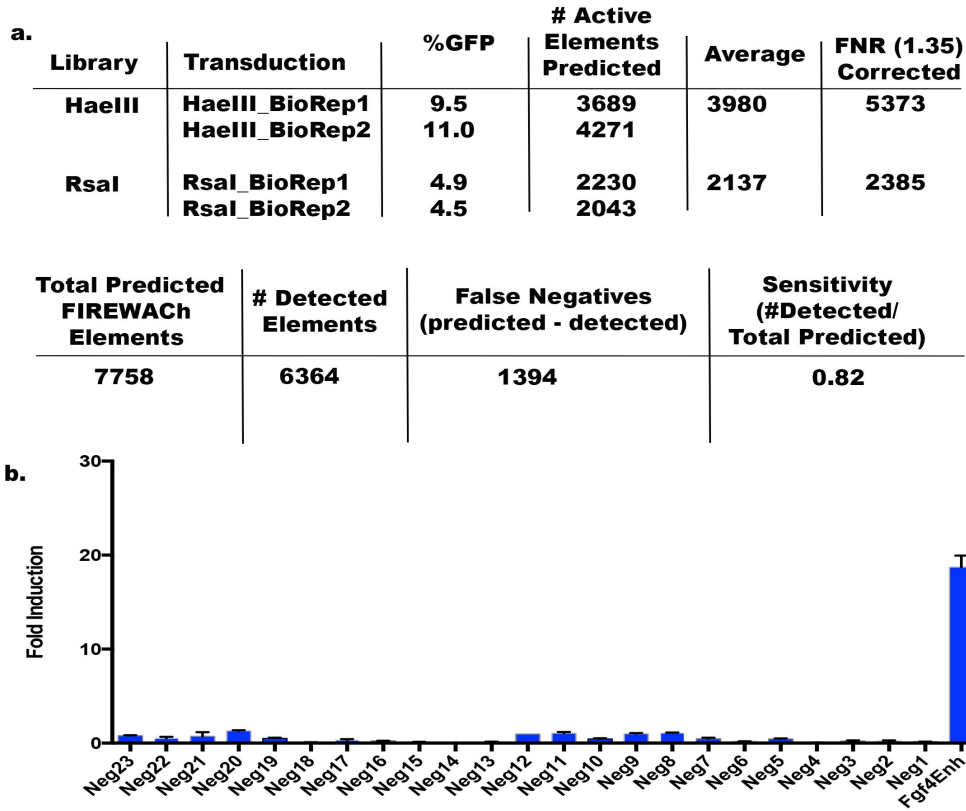
**Supplementary Figure 4. FIREWACH Selects Elements with Wide Range of Activity.** The distribution of Luciferase activities observed for reporter plasmids transfected into ESCs and containing individual elements recovered from input NFR-GFP-LV libraries **(a)** or GFP+ transduced cells **(b)** is shown. The Median level for the fold induction observed for each dataset is plotted on the right (maroon bar). In addition to exhibiting a greater percentage of active CRMs, the FIREWACH elements demonstrate a wide range of activities from 2 fold to >100 fold induction and a 10-fold greater median for luciferase activity. The horizontal line across each panel depicts the level of 2.5-fold induction.



**b.**

Activation Threshold for 'True Positive'	% Above Threshold	False Positive Rate
<b>p-value &lt;0.05</b>	<b>76</b>	<b>.2407</b>
<b>&gt; 2-fold induction</b>	<b>78</b>	<b>.2222</b>
<b>&gt; 5-fold induction</b>	<b>66</b>	<b>.3333</b>
<b>&gt; 8-fold induction</b>	<b>54</b>	<b>.4630</b>

**Supplementary Figure 5. FIREWACH Specificity.**(a) In order to determine the Specificity of FIREWACH for selecting active CRMs, transgenic NFR elements were recovered from the integrated LVs within transduced GFP+ cells and the 'true' transcriptional activation rate was determined for 53 individual elements using luciferase assays (FIREWACH Elements). Shown is the validation rate as a function of the threshold defined for luciferase activation. At a threshold of 2-fold, nearly 80% of the FIREWACH Elements are active, whereas at a threshold of 8, this is observed for approximately 50% of FIREWACH DNAs. Similar analysis of 20 DNAs recovered from integrated LVs within GFP- negative transduced ESCs (Negative Elements), 41 Input NFR-GFP-LV library (Input library NFRs), or randomly selected HaeIII- or RsaI DNA fragments generated by the digestion of murine genomic DNA (Random gDNA) shows that FIREWACH Elements display a greater percentage of active elements ('true positives') at all thresholds. Plotted is the proportion of elements that pass a given threshold and considered 'true positive'. P-value = elements that activated transcription at statistically significant level above empty vector. **(b)** The FIREWACH False Discovery rate (FDR) calculated for each threshold from the data in (a). The most likely sources of elements contributing to the FDR could include 'stowaway constructs' ie inactive transgenes integrated in a cell that also contains a transgene that is actively expressing GFP, incomplete purity of the FACS-selected cell population, or possible position effects of adjacent genomic elements interacting with an inert transgene to cause 'false' GFP activation (also see Supplementary Fig. 9).



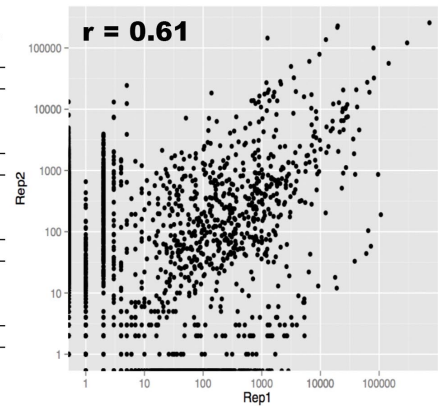
**Supplementary Figure 6. FIREWACH Sensitivity. (a)** To assess FIREWACH's success in detecting all potential active elements present in the input NFR-GFP-LV libraries, the theoretical number of expected elements (7,758) was estimated by multiplying the total number of unique NFR DNAs in each of the input NFR-GFP libraries by the percentage of transduced cells that exhibit GFP expression and multiplying by the factor 1.35 to compensate for False Negatives as determined in Supplementary Fig 2. Since FIREWACH-seq detected 6,364 elements, 1394 fewer than the estimated total active elements, the sensitivity is estimated to be 0.827. **(b)** As an additional assessment of sensitivity, the activity of 20 elements recovered from GFP- negative transduced cells were individually tested using luciferase assays in transfected ESCs. None of the elements displayed 2-fold or greater luciferase activity, demonstrating that less than 5% of these elements are active. This is less than the 16-20% active elements detected in the input NFR-GFP-LV libraries, consistent with a high level of selectivity by FIREWACH.

<b>Factors Contributing to the False Negative Rate</b>		
<b>Factor</b>	<b>Cause</b>	<b>Remedy</b>
<b>Suppression of transgene expression</b>	Integration at unfavorable genomic locus	Transduce an excess number of cells over estimated library complexity
<b>Loss of library complexity</b>	PCR bias or non-linear amplification during recovery from genomic DNA	Perform multiple independent amplification reactions using minimum number of cycles; pool samples
	Loss of cells in culture	Maintain high cell numbers with passage
	Loss of cells during FACS	Perform FACS using a large excess of cells
<b>Inactivation or absence of active transcriptional element in input library</b>	Restriction enzyme cuts at important TF binding sites, or does not target region containing functional sequences	Use multiple restriction enzymes, with distinct recognition sequences, to create the input NFR library
<b>Factor's Contributing to the False Positive Rate</b>		
<b>Factor</b>	<b>Cause</b>	<b>Remedy</b>
<b>Multiple transgenes per cell</b>	Multiplicity of Infection too high	Use low MOI determined to favor single integration events
<b>Contamination of GFP+ cells with GFP- cells</b>	Inadequate cell sorting	Ensure single cell suspension prior to FACS
		Perform multiple rounds of FACS

**Supplementary Figure 7. Factors influencing the balance of False Negative and False Positive Rates in FIREWACH.** Listed are several factors influencing this balance and corrective measures that can be taken to optimize the isolation of true positives using FIREWACH.

**a.****Table: Pearson Correlation Between Replicates**

<b>Sample:</b>		<b>TechRep1</b>	<b>TechRep2</b>	<b>TechRep3</b>
<b>HaeIII_BioRep1</b>	<b>T-Rep1</b>	1		
	<b>T-Rep2</b>	0.87	1	
	<b>T-Rep3</b>	0.98	0.86	1
<b>RsaI_BioRep1</b>	<b>T-Rep1</b>	1		
	<b>T-Rep2</b>	0.86	1	
	<b>T-Rep3</b>	0.94	0.95	1
<b>HaeIII_BioRep2</b>	<b>T-Rep1</b>	1		
	<b>T-Rep2</b>	0.98	1	
	<b>T-Rep3</b>	0.97	0.97	1
<b>RsaI_BioRep2</b>	<b>T-Rep1</b>	1		
	<b>T-Rep2</b>	0.99	1	
	<b>T-Rep3</b>	1	0.99	1
<b>Input Library (Run1 versus Run2)</b>		<b>0.84</b>		
<b>HaeIII (all) versus in silico</b>		<b>0.001</b>		
<b>RsaI (all) versus in silico</b>		<b>0.001</b>		
<b>Correlation between Biological Replicates</b>				
<b>FIREWACH-seq Rep1 vs Rep2</b>		<b>0.61</b>		

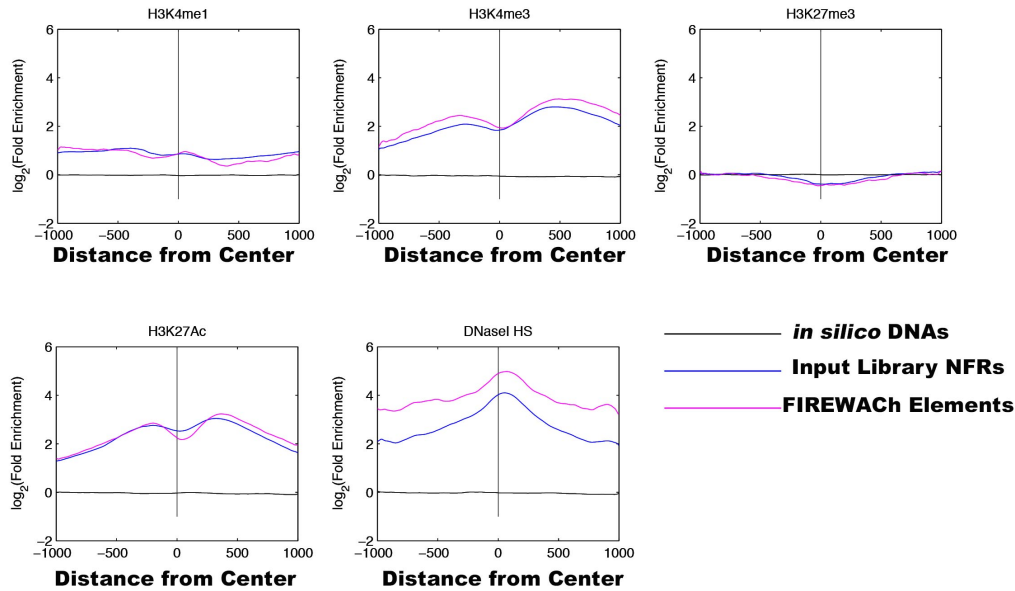
**b.****Supplemental Figure 8: Correlation between technical and biological replicates.**

A) Table of correlation coefficients between technical replicates. To measure reproducibility of our PCR rescue and sequencing strategy, each sample (i.e. GFP<sup>+</sup> cells from a single transduction) was assayed independently three times as technical replicates. Technical replicates consist of independent PCR amplification of selected NFRs, library preparation, and sequencing from a common template (e.g. GFP<sup>+</sup> cells transduced with HaeIII NFR-GFP LV). The correlation between technical replicates was measured pairwise by binning the genome into windows of 100bp and computing the Pearson correlation of the genomic coverage between all pairs of the coverage vectors, which represent our sequencing dataset. B.) Scatterplot of correlation between biological replicates. To generate the scatterplot, the genome was binned into windows of 100bp and the genomic coverage was estimated for each genomic bin. The coverage values of the two replicate experiments are shown in the scatterplot. The scatterplot was done using the R statistical software (<http://www.r-project.org/>). A biological replicate represents the results of the three independent technical replicate sequencings of a single HaeIII and RsaI transduction combined. The two transductions are then compared (Rep1 vs. Rep2).

**NOTE:** Variable transduction efficiency is likely to play a large role in reducing the overall correlation between biological replicates, ( $r=0.61$ ). We interpret this as an indication that a large number of elements were found in one sample and not the other, as seen by appearance of data points along the axes. Another potential source effecting reproducibility is PCR bias -i.e. the under or over-representation of PCR products, typically due to stochastic amplification in the initial rounds of amplification. Therefore, any analysis utilizing read count, such as measuring the correlation coefficient between two replicates, will necessarily be depressed as it contains the stochastic bias within the samples themselves. Thus, although a general trend showing the relatedness of these samples is observed in the scatterplot (panel b), the somewhat low extent of this correlation reflects both PCR bias as well as whether particular elements are identified in both replicates. This highlights our conclusion that Selectivity and Sensitivity are better

measures of reproducibility of FIREWACH's performance rather than analyses involving a comparison of read counts of elements recovered from Biological replicates.

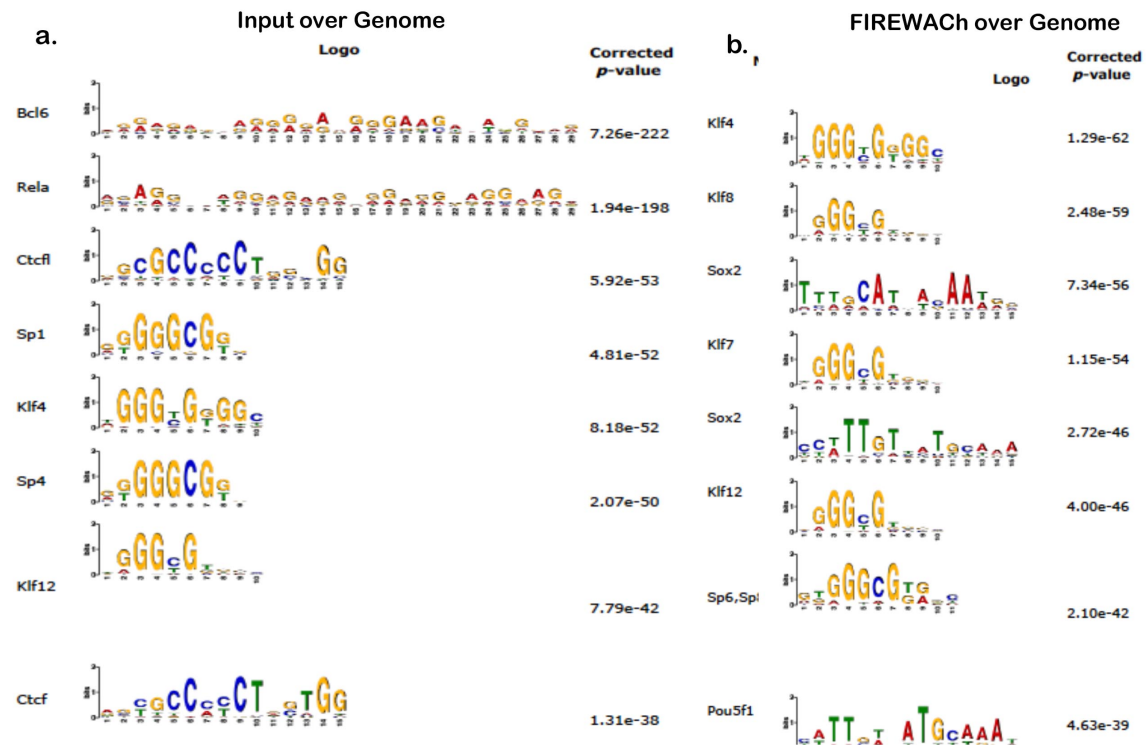
**a.**



**b.**

	H3K4me1	H3K4me2	H3K27me3	H3K27Ac	DNaseI HS
<b>Input vs <i>in silico</i></b>	$10^{-211}$	$<10^{-320}$	$3.5 \times 10^{-4}$	$<10^{-320}$	$<10^{-320}$
<b>FIREWACH vs. <i>in silico</i></b>	$10^{-213}$	$<10^{-320}$	$1.6 \times 10^{-9}$	$<10^{-320}$	$<10^{-320}$
<b>FIREWACH vs Input</b>	$2.6 \times 10^{-1}$	$1 \times 10^{-22}$	$1.6 \times 10^{-2}$	$2.9 \times 10^{-5}$	$10^{-217}$

**Supplementary Figure 9. Comparison of Chromatin Marks Over Random, Library, and FIREWACH Elements.** (a) Read-density profiles of H3K4me1, H3K4me3, H3K27me3, H3K27Ac and DNaseI HS loci for 1kb+/- genomic regions associated with input Library NFR DNAs, FIREWACH Elements, or *in silico*- generated HaeIII and RsaI genomic DNA fragments. Y axis= Read density for each of the features indicated at the top of each panel (calibrated such that the signal density is normalized with respect to the average signal density in the window positioned at 15kb-20kb upstream of the FIREWACH-seq read); X-axis = genomic locations relative to the DNA fragments. To aid visualization, the signal is smoothened using a 150bp moving average and scaled using a logarithm transform. Note that all features associated with active transcriptional elements (H3K4me1, H3K4me3, H3K27Ac and DNaseI HS) are enriched in the input library and FIREWACH-seq data sets compared to the *in silico* DNAs, whereas the opposite is true for the H3K27me3 mark associated with transcriptional silencing. This data is summarized in (b) along with p-values obtained using the Kolmogorov-Smirnov 2-sample test (All p-values are /2 to account for two-sided test).



**Supplementary Figure 10. Motifs Enriched in FIREWACH DNAs.** The top motifs generated by the analysis of DNA sequences of Distal elements from the input NFR-GFP-LV library **(a)** or distal FIREWACH elements **(b)**. We also analyzed the FIREWACH distal elements using library NFRs as background (Supplementary Table 7). Motif analysis was performed as detailed in the Methods and the complete list of enriched motifs, associated  $p$  values, and PWMs for the discussed motifs are presented in Supplementary Table 7.

## References

1. Ying, Q.-L. *et al.* The ground state of embryonic stem cell self-renewal. *Nature* **453**, 519–523 (2008).
2. Murtha, M., Wang, Y., Basilico, C. & Dailey, L. Isolation and analysis of DNA derived from nucleosome-free regions. *Methods Mol. Biol.* **977**, 35–51 (2013).
3. Ren, B. & Dynlacht, B. D. Use of chromatin immunoprecipitation assays in genome-wide location analysis of mammalian transcription factors. *Meth. Enzymol.* **376**, 304–315 (2004).
4. Lois, C., Hong, E. J., Pease, S., Brown, E. J. & Baltimore, D. Germline transmission and tissue-specific expression of transgenes delivered by lentiviral vectors. *Science* **295**, 868–872 (2002).
5. Yaragatti, M., Basilico, C. & Dailey, L. Identification of active transcriptional regulatory modules by the functional assay of DNA from nucleosome-free regions. *Genome Research* **18**, 930–938 (2008).
6. Charrier, S. *et al.* Quantification of lentiviral vector copy numbers in individual hematopoietic colony-forming cells shows vector dose-dependent effects on the frequency and level of transduction. *Gene Ther.* **18**, 479–487 (2011).
7. Caporaso, J. G. *et al.* Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci. U.S.A.* **108** Suppl 1, 4516–4522 (2011).
8. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
9. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
10. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**, R25 (2009).
11. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**, 562–578 (2012).
12. Marks, H. *et al.* The Transcriptional and Epigenomic Foundations of Ground State Pluripotency. *Cell* **149**, 590–604 (2012).
13. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).
14. Kruskal, W. H. & Wallis, W. A. Use of ranks in one-criterion variance analysis. *Journal of the American statistical ...* (1952).
15. Gao, Z. *et al.* PCGF homologs, CBX proteins, and RYBP define functionally distinct PRC1 family complexes. *Mol. Cell* **45**, 344–356 (2012).
16. Zhang, J. *et al.* SFMBT1 functions with LSD1 to regulate expression of canonical histone genes and chromatin-related factors. *Genes Dev.* **27**, 749–766 (2013).
17. Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research* **37**, W202–8 (2009).
18. Machanick, P. & Bailey, T. L. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* **27**, 1696–1697 (2011).
19. Bulyk, M. L., Huang, X., Choo, Y. & Church, G. M. Exploring the DNA-binding

- specificities of zinc fingers with DNA microarrays. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 7158–7163 (2001).
20. Berger, M. F. *et al.* Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell* **133**, 1266–1276 (2008).
  21. Jolma, A. *et al.* DNA-binding specificities of human transcription factors. *Cell* **152**, 327–339 (2013).
  22. Mathelier, A. *et al.* JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Research* (2013). doi:10.1093/nar/gkt997
  23. Matys, V. *et al.* TRANSFAC and its module TRANSCOMP: transcriptional gene regulation in eukaryotes. *Nucleic Acids Research* **34**, D108–10 (2006).
  24. Yuan, H., Corbi, N., Basilico, C. & Dailey, L. Developmental-specific activity of the FGF-4 enhancer requires the synergistic action of Sox2 and Oct-3. *Genes Dev.* **9**, 2635–2645 (1995).
  25. Akhtar, W. *et al.* Chromatin position effects assayed by thousands of reporters integrated in parallel. *Cell* **154**, 914–927 (2013).